HIT Standards Committee –Vocabulary Task Force
Panel 4: Terminology Services Vendors, Developers, Implementers
September 2, 2010
Testimony submitted by Ken Buetow, Ph.D.
Director, Center for Bioinformatics and Information Technology
National Cancer Institute, National Institutes of Health

The Task Force has created a comprehensive collection of questions seeking input on the requirements and priorities for vocabulary infrastructure that would provide an ideal "one stop shop" for value sets, convenience subsets, and entire vocabularies needed to achieve Meaningful Use. To that end, the NCI is responding to these questions -- drawing on our collective experiences in creating, publishing, and consuming terminologies and value sets, as well as in developing the tools and infrastructure involved in operationalizing and supporting these terminology-focused capabilities in the cancer community and beyond.

1. **What are the requirements for a centralized infrastructure to implement "one-stop shopping" for obtaining value sets, subsets, and vocabularies for meaningful use?**

2. **Which requirements or functionalities are urgent, i.e., absolutely required to support "meaningful use"? Which would be most useful immediately? What would be a staged approach over time to get to the desired end state?**

3. **Is there a difference between versioning for clinical documentation vs. versioning for reported measures, i.e., when do you go live with a change in the EHR vs. when do you use the new version for measures?**

4. **How do you manage versioning in clinical decision support vs. changes in value sets?**

There are several dimensions we consider critical to the discussion:

- In addition to terminology distribution, one should pay equal attention to terminology authoring, in order to achieve a scalable and effective process in terminology support for Meaningful Use.
- In particular, centralized serving and authoring of terminology should be weighed against a federated mechanism, to arrive at a deployment topology perspective most suited to the agility and responsiveness in a clinical context required for Meaningful Use.
- Domain federation as a source of terminology should be compared against having a central source of terminology, in determining which strategy would best accommodate the terminology needs for Meaningful Use. In this model, domains-of-interest define the content specific to their domain so that it can then be integrated into the larger compendium of cross-domain semantics via a centralized process, thereby maximizing the relevance of domain-specific terminologies in domain-specific contexts.
- Support for localizations, in the form of constraints as well as extensions, is also critical.
- In this regard, a layered semantic space should be considered against having a "flat" file of terminologies or value sets. We will expand on these dimensions in greater detail below.

As part of the caBIG® program, NCI uses both terminologies and value sets to develop and describe information models, data elements, and data types to:

- Create clinical case report forms (CRFs) for all NCI Cooperative Group Trials, as well as to author large numbers of CRFs for trials conducted by NCI-designated Cancer Centers and affiliated groups. These forms are programmatically loaded into Clinical Data Management Systems and utilized for both trial design and execution. Structured data collected during the course of a clinical trial is then available for analysis on both an intra- and inter-trial basis.
- Define the metadata for public interfaces to information systems to enable semantic interoperability among different modalities in basic and clinical research in order to advance knowledge of oncologic processes and their treatment.
- Provide a framework and a foundational basis for other agencies and organizations -- including the U.S. Food and Drug Administration (FDA), the Clinical Data Interchange Standards Consortium (CDISC), and the National Council of Prescription Drug Providers (NCPDP) -- to create, publish and maintain their terminology subsets and metadata in support of a multitude of purposes, including regulatory reporting (e.g. Structured Product Labeling and Device Event Problem Codes), FDA submissions (e.g. Study Data Tabulational Model, SDTM), and pharmacy communications (SCRIPT 10.5).

These uses impose the following requirements for the vocabulary infrastructure developed at NCI [1]:

- High uptime and full redundancy, as with any key production system. Ideally, content should flow to localized repositories in the way that Domain Name Service (DNS) flows from a primary sort of truth to local systems that maintain synchronized copies for better performance.
- Rapid turnaround for new or modified concepts in vocabularies and additions and changes to value lists (24 hours or less).
- Acceptance of multiple vocabularies for specific purposes rather than a single vocabulary that supports all research in all fields.
- Clear and easy mechanism for community input with regard to vocabulary content. Ideally, vocabulary and metadata should be developed by a community (ensures greater uptake) while being managed by coordinating groups.
- Staff support for maintaining vocabularies and data elements and their associated value lists. Such staff should be experts in terminology and metadata with some amount of expertise in biomedicine. It is possible to separate the terminology/metadata expertise from subject matter expertise, but this requires careful coordination, particularly if the subject matter experts are volunteer community members.
- Multiple versions of a standard maintained. Multiple versions, e.g. of AJCC staging criteria, or MedDRA, are needed to support treatment plans over time (e.g. relapse of a cancer after 10 years) and to support users who update the terminologies and value sets at different times. Clinicians may not know which version was used in the past, so access by version date is important.
- A notification system for standard vocabulary and value set updates.

We believe that all of the above requirements are important to support Meaningful Use. In particular, there are two aspects we have recognized as most useful in our experience in terminology development:

- The NCI vocabulary infrastructure supports a federated mechanism for distribution and authoring of terminologies, subsets and value sets, and a collaborative authoring framework with corresponding tools and a logic-based governance approach [2]. The goal is to mitigate the risk of occurrence of a common dilemma in terminology development that results from a conflict between agility/rapid turnaround of new concepts critical to use cases such as report measures, and governed stability and continuity of a terminology critical to use cases such as clinical documentation, electronic health records and clinical decision support. The NCI approach of federated terminology development successfully supports both types of use cases for our

community, in efforts such as the Common Terminology Criteria for Adverse Events (CTCAE), and the Cancer Clinical Information Suite, an electronic health record framework for oncology.

- Based on our experience over the past six years with a relatively "flat" approach to terminology management, NCI terminology development is now being based on a "layered" approach involving information, data type, and terminology models all supported and inter-linked with an overarching ontology. This layered approach of defining semantics enables localization of terminology or metadata to be defined in relation to the standard in terms of constraint or extension, and mitigates the potential conflict between need for localization support vs. need for centralized data submission and integration. [3]

5. Where are you using value sets and subsets? For what domains? How many value sets and subsets?

More than a hundred named, tagged NCI Thesaurus (NCIt) subsets and value sets with some 20,000 values are maintained in collaboration with a variety of partners, most notably:

- U.S. Food and Drug Administration (FDA): Many FDA subsets are maintained in NCIt and required for regulatory reporting and other purposes. These include: 16 subsets used by Structured Product Labeling (SPL) for submission of proposed labeling by all manufacturers using electronic formats; Device Event Problem Codes subsets used for the reporting of medical device problems to FDA (roughly 3,000 different reporting locations used these subsets in 2009); and Individual Case Safety Report (ICSR) subsets used for adverse event reporting (proposed regulations for electronic submissions will create similar levels of use for these subsets).
- Clinical Data Interchange Standards Consortium (CDISC): All CDISC controlled terminology is maintained and published as NCIt subsets, including the Study Data Tabulation Model (SDTM), an approved standard for FDA submissions that has been downloaded more than 14,000 times in over 60 countries, primarily for institutional use.
- National Council of Prescription Drug Providers (NCPDP): Three NCIt subsets have been adopted as part of the SCRIPT (10.5) and Telecommunication (D.3) standards employed by some 200 vendors serving approximately 15,000 pharmacies nationwide.

NCIt subsets are used extensively within NCI and caBIG® systems. The extensive commitment that NCI has made to create active collaborations uniting large segments of the cancer and biomedical community has facilitated widespread adoption and reuse of value sets and code sets. For example, community-based resources such as the caBIG® Knowledge Centers and the various caBIG® subject matter workspaces are forums in which patients, researchers and clinical care providers interact to identify the need for value sets and subsets, as well as working together on drafting, review, deployment and maintenance. Community-based resources such as the caBIG® Knowledge Centers and Support Service Providers supplement and extend the ability of the NCI to encourage adoption and proper use of subsets and value sets across the community.

NCI metadata services are one of the primary means by which these and other vocabulary subsets and value sets are used in NCI and other systems and applications. There are over 135 information models represented as ISO 11179 metadata in caDSR, recording the data semantics in software applications and in services on NCI's caGRID. This metadata encompasses some 20,000 data elements referencing subsets and value sets drawn from NCI terminology services, many shared between users. Significant users of metadata-based subsets and value sets are:

- NCI internal programs: Many NCI divisions, centers, and programs make use of metadata-defined subsets and value sets. The Cancer Therapy Evaluation Program (CTEP) alone uses more than 10,000.

- Other NIH institutes: Some 5,000 data elements are used by the National Heart, Lung and Blood Institute (NHLBI), the National Institute of Child Health and Development (NICHD), and the National Institute of Dental and Craniofacial Research (NIDCR).
- caBIG®: Virtually all projects use metadata-supported subsets and value sets in their models, interfaces and information content.
- Biomedical Research Integrated Domain Group (BRIDG): Over 1,600 data elements are used in this domain analysis model of clinical and pre-clinical protocol-driven research created in collaboration with the Clinical Data Interchange Standards Consortium (CDISC), the HL7 Regulated Clinical Research Information Management Technical Committee (RCRIM TC), and the U.S. Food and Drug Administration (FDA).
- CancerGrid: UK deployments of NCI terminology and metadata services are supporting a variety of projects at both the UK and EU levels.

6. In your experience with creating, disseminating, updating and/or using value sets, subsets, and entire vocabularies, what works and what does not work?

7. What functions are required that users have not yet appreciated?

NCI has devoted many years of effort to standardizing and sharing its vocabulary subsets and value sets, as an integral part of efforts to build comprehensive content and technology standards that can create interoperability and synergies in our clinical, research, and public health systems. We have learned many lessons and worked to create best practices, involving process and institutional lessons at least as much as technical ones:

- Work closely with stakeholders to identify content, operational, and technical requirements. Existing terminologies, subsets and value sets reflect specific purposes and use cases, including important institutional, professional and regulatory constraints. For example, MedDRA is the internationally accepted standard for adverse event reporting, and NCI (with extensive community input) redesigned its CTCAE adverse event terminology to be a harmonized intersection with a subset of MedDRA, even while recognizing MedDRA's limitations and preferring other terminologies for most coding of diagnoses and findings.
- Subsets and value sets, as well as terminologies, terminology maps and extensions generally, are built for specific operational uses. It is important that information be provided about the purpose of a subset or value set, its provenance, currency, intellectual property limits, and other information pertaining to suitability for use, ideally as an organic part of the subset or value set.
- Especially with regulatory terminology or terminology that is to be adopted and used by large diverse communities, such as regulated device manufacturers or pharmacies, technical documentation about the structure and proper deployment of subsets and value sets must be provided. In many cases, online or telephone support for adopters is required to ensure proper use.
- Not all subsets and value sets are intended for use by regulated or large groups; some are conveniences intended for the use of a few people or perhaps a pair of cooperating groups. In the past, support for such small-scale uses has had to be limited, but with the advent of interactive services such as the NCI's new Common Terminology Services (CTS2), such uses are expected to grow rapidly. On-line training and end user support for such users will be an important role for organizations providing terminology services.
- The infrastructure needed to ensure best practices in the creation, maintenance, and distribution of vocabulary subsets and value sets is complex and evolving. Some centralization is important to cost-effective provision of high quality services. At the same time, the vital role of interaction with both creators and users seems likely to put limits on centralization given current technologies.

- Use of available content and technical standards, and contributing to the development of such standards, is important to both implementation and adoption of subsets and value sets.
- Vocabulary subsets and value sets need robust and transparent mechanisms for input by affected communities. This will mean providing methods for public input for publicly defined sets.
- Workflow support is vital to the creation and maintenance process for any robust set of vocabulary subsets and value sets.
- Distribution formats often need to cover a very broad spectrum of users and implementations, ranging from very simple text files of terms and codes through to complex representations of full underlying vocabulary data. Failure to analyze and meet the specific requirements of target user communities can greatly impair use.
- The pace and patterns of versioning vary enormously with the purpose. Cutting edge activities often require 24-hour turnaround to support coding needs, while in other areas it is important to establish much longer revision cycles. History, change tracking, and version labeling and description require a uniform framework but also sufficient flexibility to address the full range of requirements.
- High quality, usable content requires an unusual blend of domain and technical expertise. Sometimes a significant part of this can come from outside contributors, but both will be required of maintenance organizations to ensure quality and consistency.
- Diversity in underlying source terminologies is a fact of life for the foreseeable future. Improvements in mapping between them, and between subsets and value sets based on them, will be crucial to interoperable health information in the years to come.

8. What human resources does it take to implement and manage value sets, subsets, and entire vocabularies? Informaticists? Clinicians? IT people? How are you organized?

9. What is your maintenance process? How do you manage updates?

Creation and revision of vocabulary subsets and value sets starts with stakeholder input and feedback. The caBIG® program has provided an important framework for broadening NCI's internal processes to include Cancer Centers, Cooperative Groups, and a broad range of other government, academic, professional and private stakeholders. Many NCI-maintained subsets and value sets involve intensive ongoing interaction with specific partners such as FDA, CDISC, and NCPDP.

Expert curators compare requests and requirements with existing subsets, value sets, and other content in EVS terminologies and, when appropriate, the metadata repository. New or revised sets are often circulated for internal and external review before implementation, but requests cleanly aligned with existing content and policies can sometimes be released after basic internal QA.

Internal QA involves a variety of human and computer processes, as describe in a recent publication (The NCI Thesaurus Quality Assurance Life Cycle. *Journal of Biomedical Informatics* 2009 June;42(3):530-539.)

Given the extensive use of NCI subsets and value sets, we often get user feedback on possible extensions or modifications. NCI then consults partners and known stakeholders in the existing sets, and will either design and implement agreed-upon changes or suggest other approaches (possibly including a separate extended or new subset or value set) to meet user needs.

Distribution is also a vital part of any process dealing with vocabulary subsets and value sets. While it is important that all such sets be accessible in file, API and browsable forms standardized for all sets,

individual maintainers and users often have their own distinctive requirements that are crucial to the success of the set.

The FDA subsets are an instructive example of matching process to specific use cases. NCI support of FDA terminology is conducted under a Memorandum of Understanding. Extension or modification of the underlying terminology is driven by FDA regulatory requirements, as well as ongoing quality assurance operations within NCI Enterprise Vocabulary Services (EVS). Subsets are defined to meet FDA operational needs, and are then reviewed and revised jointly by FDA, NCI, and other stakeholders during pre-release development. Distribution mechanisms are tailored to meet FDA and regulated industry requirements, as are online support and end user documentation. Publication and distribution of the subsets are mission critical responsibilities of NCI EVS and technical operations groups. Post-release revision of the subsets is an ongoing responsibility of the NCI EVS content curation group. The revisions are driven by the needs of the FDA and regulated industry, which surface their needs to the FDA and NCI project officers who oversee the revision and release process. Broadly similar arrangements are in place for the CDISC and NCPDP subsets.

A recent example of broad-based community collaborative development of a heavily used value set is the 4th edition of the Common Terminology Criteria for Adverse Events (see https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/CTCAE), which was produced using the LexWiki service. The CTCAE v4 effort was sponsored by the caBIG® Vocabulary and Common Data Elements Workspace (VCDE) (see https://wiki.nci.nih.gov/display/VCDE/VCDE+Wiki+Home+Page), which conducts an ongoing program to review and certify terminological products for use in caBIG® and establishes the criteria the community will follow to achieve and sustain semantic consistency across clinical and research resources. Community-based resources such as the caBIG® VCDE, Knowledge Centers and Support Service Providers supplement and extend NCI's ability to encourage adoption across the community and support proper use of terminology and metadata artifacts.

NCI's metadata environment has formal governance mechanisms that provide direction and oversight of the creation, deployment and re-use of Value Domains, Valid Value Lists and other components of the Common Data Elements that compose the bulk of NCI's metadata. Value sets and subsets drawn from the terminology services are bound within the metadata domain to specific business uses and are provided with situation-specific contextual representations. These business-specific representations are critical to understanding the nuanced meaning of data described by the metadata, and so have received careful curation from the beginning of the services.

10. **What national resources and services could be leveraged to reduce the level of effort required for local implementations? What is the irreducible minimum of local work at an implementation site, or within an organization or system?**
11. **How do you manage distribution of updates with local variations and optionality? Unique subsets? Local mappings?**

Beginning in the Fall of 2010, NCI terminology services will be extended to fully support inter-terminology mapping, local extensions to standard terminologies, and value sets, as called for by HL7 and OMG standards. Mappings have been identified as critical, and versioning and updating mappings regularly is critical; but it also should be noted that one terminology cannot always be swapped out for another cleanly. Mappings between local terminologies and standard terminologies, defining local extensions to standard terminologies, and identifying appropriate value sets to leverage, are among the irreducible minimum of local work at an implementation site. The NCI terminology services software (including terminology browsers being modified to browse value sets) is open source, and can conveniently be used as-is or modified to support local implementations.

12. What metadata do you maintain and how do you maintain versioning?

13. How does an application know which value set is for which purpose? How is the specific context for a value set maintained at the message data element level of specificity? How is the English language intent of the value set context documented and maintained?

Value set support under the HL7 CTS2-compliant services available from NCI will include services to create, manage, organize, search and retrieve value set specifications and to record extensive metadata about them, including the reason the set was created, its intended purpose, responsible parties and their contact information, version history, and terminologies and terminology versions from which the value set is drawn.

14. What are lessons learned about web links vs. storage of the vocabulary or other artifact in a physical repository?

The NCI terminology infrastructure supports web links as well as allows users to deploy a local physical repository using open-source software we provide. A local physical repository can potentially mitigate against performance or availability issues compared to using web links to a centralized repository. We do not have a best practice currently regarding this.

15. How do you manage distribution of updates to multiple sites?

NCI supports a notification system for standard vocabulary and value set updates. We do not provide a "push" mechanism for updates at this time.

16. Where is local customization appropriate and how much customization is acceptable?

17. What has to be local in an EHR implementation vs. what can be external in a vocabulary repository?

This again involves a potential conflict between need for localization support vs. need for centralized data submission and integration. NCI mitigates this by adopting a "layered" approach to terminology development, as described in the answer to Questions 1-4. Further, the NCI terminology infrastructure allows even information such as "pick lists" corresponding to value sets for an EHR implementation, to be recorded externally in a vocabulary repository, further enhancing the capability for subsequent centralized data integration.

[1] James J. Cimino. "Desiderata for Controlled Medical Vocabularies in the Twenty-First Century." Methods of Information in Medicine 1998 Nov; 37(4-5): 394-403.

[2] Keith Eugene Campbell. "Distributed Development of a Logic-Based Controlled Medical Terminology." PhD Thesis, Stanford University, June 1997.

[3] HL7 Version 3 Standard: Refinement, Constraint and Localization to Version 3 Messages, Release 2. 8/20/2007.